Introduction to Hyper-Threading Technology

Doucment Number: 250008-002

2001

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an ordering number and are referenced in this document, or other Intel literature, may be obtained from:

Intel Corporation
P.O. Box 7641
Mt. Prospect IL 60056-7641

or call 1-800-879-4683
or visit Intel's website at http:\\www.intel.com

PAGE

ABOUT THIS PAPER

This paper describes the new Hyper-Threading technology from Intel®. Hyper-Threading technology is a new feature in the IA-32 Intel Architecture that provides a performance boost for future Intel IA-32 processors based on the Intel® NetBurst™ microarchitecture. Included in this paper is a brief overview of Hyper-Threading technology and description of how it increases the performance of operating-system and application software written to run on IA-32 processors.

INTRODUCTION TO HYPER-THREADING TECHNOLOGY

Throughout the evolution of the IA-32 Intel Architecture, Intel has continuously added innovations to the architecture to improve processor performance and to address specific needs of compute-intensive applications. The latest of these innovations is Hyper-Threading technology, which Intel has developed to improve the performance of IA-32 processors when executing multiple-processor (MP) capable operating systems and multi-threaded applications.

1.0 HYPER-THREADING TECHNOLOGY OVERVIEW

Hyper-Threading technology enables a single physical processor to execute two separate code streams (called *threads*) concurrently. Architecturally, an IA-32 processor with Hyper-Threading technology consists of two *logical processors* (see Figure 1), each of which has its own IA-32 architectural state[1]. After power up and initialization, each logical processor can be individually halted, interrupted, or directed to execute a specified thread, independently from the other logical processor on the chip. Unlike a traditional dual processor (DP) configuration that uses two separate physical IA-32 processors (such as two Intel® Xeon™ processors), the logical processors in a processor with Hyper-Threading technology share the execution resources of the processor core, which include the execution engine, the caches, the system bus interface, and the firmware.

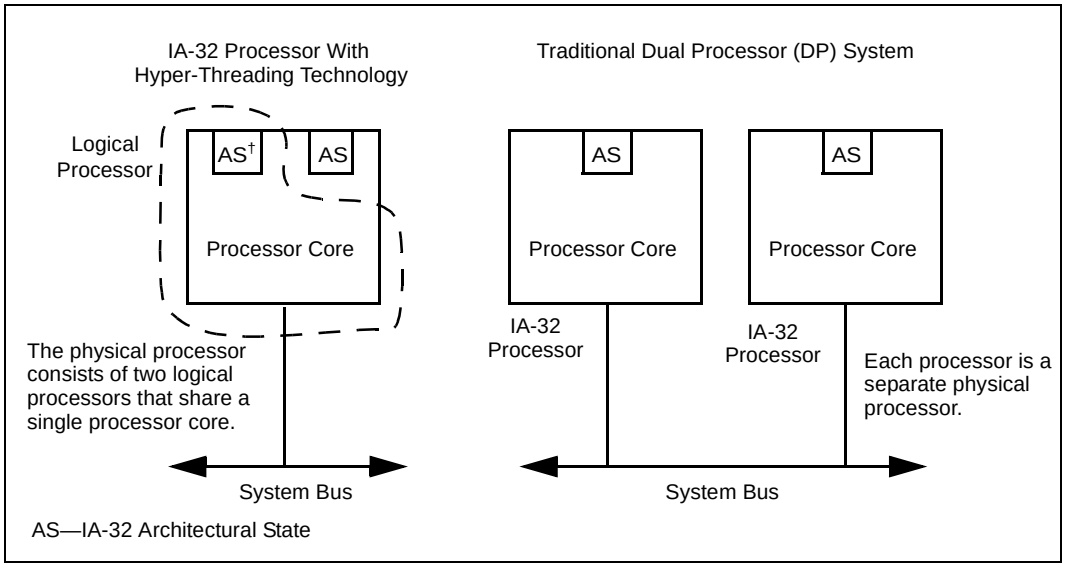

Figure 1-1. Comparison of an IA-32 Processor with Hyper-Threading Technology and a Traditional Dual Processor System.

Hyper-Threading technology is designed to improve the performance of IA-32 processors by exploiting the multi-threaded nature of contemporary operating systems, server applications, and workstation applications in such a way as to increase the use of the on-chip execution resources available in the Intel NetBurst microarchitecture.

Virtually all contemporary operating systems (including Microsoft Windows* and Linux*) divide their work load up into processes and threads that can be independently scheduled and dispatched to

1. The architectural state that is duplicated for each logical processor consists of the IA-32 data registers, segment registers, control registers, debug registers, and most of the MSRs. Each logical processor also has its own advanced programmable interrupt controller (APIC).

run on a processor. The same division of work load can be found in many high-performance applications such as database engines, scientific computation programs, engineering-workstation tools, and multi-media programs. To gain access to increased processing power, most contemporary operating systems and applications are also designed to execute in DP or MP environments, where, through the use of symmetric multiprocessing (SMP), processes and threads can be dispatched to run on a pool of processors.

Hyper-Threading technology leverages the process- and thread-level parallelism found in contemporary operating systems and high-performance applications by implementing two logical processors on a single chip. This configuration allows a thread[2] to be executed on each logical processor. Instructions from both threads are simultaneously dispatched for execution by the processor core. The processor core executes these two threads concurrently, using out-of-order instruction scheduling to keep as many of its execution units as possible busy during each clock cycle.

The increase in instruction processing throughput that Hyper-Threading technology provides results from a combination of two things:

- The design of the Intel NetBurst micro-architecture
- The mix of IA-32 instructions typically found in multi-threaded code.

The Intel NetBurst micro-architecture has been designed to provide optimum performance when executing a single instruction stream (a single thread of execution); however, typically (even with highly optimized code) not all of the available execution units are used during each clock cycle. On average, when executing code that uses a typical mix of IA-32 instructions, only 35%[3] of the execution resources of the Intel NetBurst micro-architecture are used. To make more efficient use of these execution resources, Hyper-Threading technology takes advantage of the inherent parallelism of multi-threaded code to provide the processor core with a second thread of execution. The two resulting threads supply the instruction scheduler with a pool of instructions that contain less interdependencies between instructions and thus more opportunities to use the processor core's available execution resources. The net result is an increase in the instruction processing throughput for the physical processor when executing multi-threaded code.

2.0 SUPPORTING IA-32 PROCESSORS WITH HYPER-THREADING TECHNOLOGY

An IA-32 processor with Hyper-Threading technology will appear to software as two independent IA-32 processors, similar to two physical processors in a traditional DP platform. This configuration allows operating system and application software that is already designed to run on a traditional DP or MP system to run unmodified on a platform that uses one or more IA-32 processors with Hyper-Threading technology. Here, the multiple threads that would be dispatched to two or more physical processors are now dispatched to the logical processors in one or more IA-32 processors with Hyper-Threading technology.

At the firmware (BIOS) level, the basic procedures to initialize multiple processors with Hyper-Threading technology in an MP platform resemble closely those for a traditional MP platform[4]. An operating system designed to run on an traditional DP or MP platform can use the CPUID instruction to detect the presence of IA-32 processors with Hyper-Threading technology. The same mechanisms that are described in the Multiprocessor Specification Version 1.4 to wake physical processors apply to the logical processors in an IA-32 processor with Hyper-Threading technology.

Although existing operating system and application code will run correctly on a processor with Hyper-Threading technology, some relatively simple code modifications are recommended to get the optimum benefit from Hyper-Threading technology.

---

2. In the remainder of this paper, the term "thread" will be used as a general term for the terms "process" and "thread."
3. This figure was obtained from Intel laboratory measurements.
4. Some relatively simple enhancements to the MP initialization algorithm are needed.

3.0 PERFORMANCE GAINS WITH HYPER-THREADING TECHNOLOGY

A processor with Hyper-Threading technology can provide a performance gain of up to 30% when executing multi-threaded operating system and application code over that of a comparable IA-32 processor without Hyper-Threading technology. When placed in MP systems, this increase in computing power will generally scale linearly as the number of physical processors in a system is increased; although as in any MP system, the scaleability of performance is highly dependent on the nature of the application.

4.0 SUMMARY

Hyper-Threading technology represents a new approach to improving the instruction throughput of processors that are targeted for servers and high-performance workstations. It also provides a view into the future of microprocessor design where the performance of a processor when executing a specific type of application or the space and power requirements of a physical processor within a server may be as important as its raw processing speed.